Data Science Capstone

Jason Miller

1/18/2024

Executive Summary

- Data collection from the public SpaceX API and the SpaceX Wikipedia page. Created a column labeled ‘class’ to classify successful landings. Data exploration tools included SQL, visualizations, Folium maps, and dashboards, with relevant columns selected as features and all categorical variables to binary using one-hot encoding. Standardized the data and utilized GridSearchCV to find the best parameters for the machine learning models
- Four machine learning models were developed: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. All models produced similar results, achieving an accuracy rate of approximately 83.33%
- All models tended to overpredict successful landings. Therefore, recommend additional data is needed to improve the accuracy of the machine learning model determination

Introduction

Falcon 9 rocket at Cape Canaveral's Complex 40 launch pad – SpaceX

Background:

- Commercial Space Age is Here
- SpaceX best pricing ($62 million vs. $165 million)
- Largely due to the ability to recover rocket (Stage 1)
- SpaceY wants to compete with SpaceX

Problem:

- SpaceY desires to train a machine learning model to predict successful Stage 1 recovery

Methodology

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Methodology

OVERVIEW OF DATA COLLECTION, WRANGLING, VISUALIZATION, DASHBOARD, AND MODEL METHODS

Data Collection Overview

Data collection process involved a combination of API requests from SpaceX's public API and web scraping data from a table in Space X's Wikipedia entry.

The next slide will show the flowchart of data collection from API, and the one after will show the flowchart of data collection from web scraping.

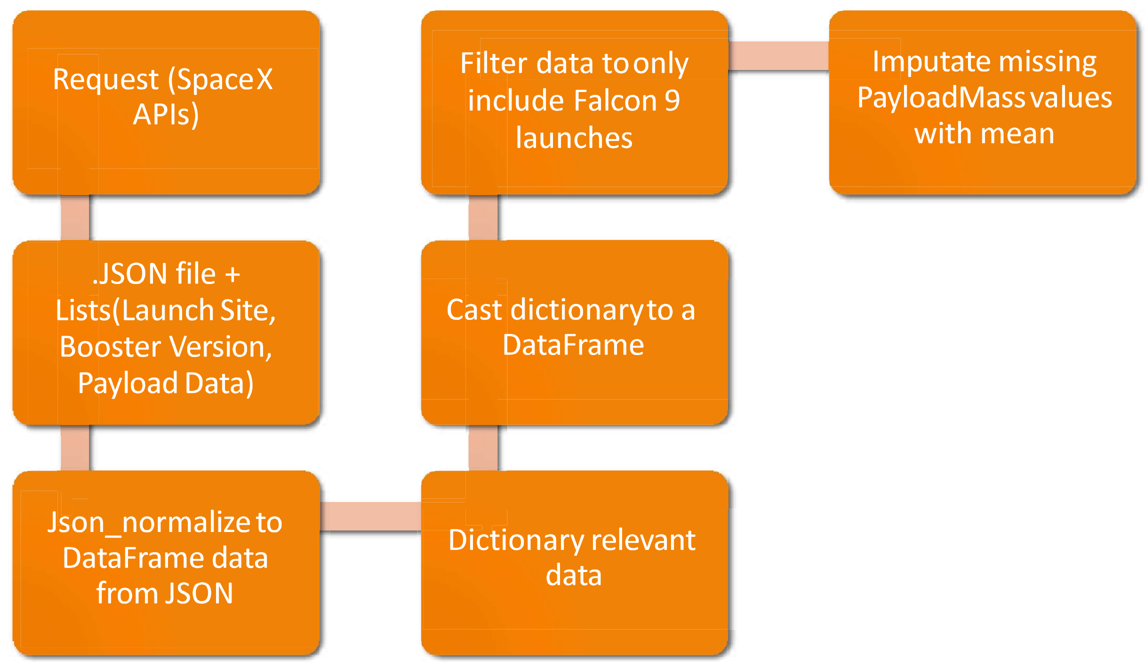
Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

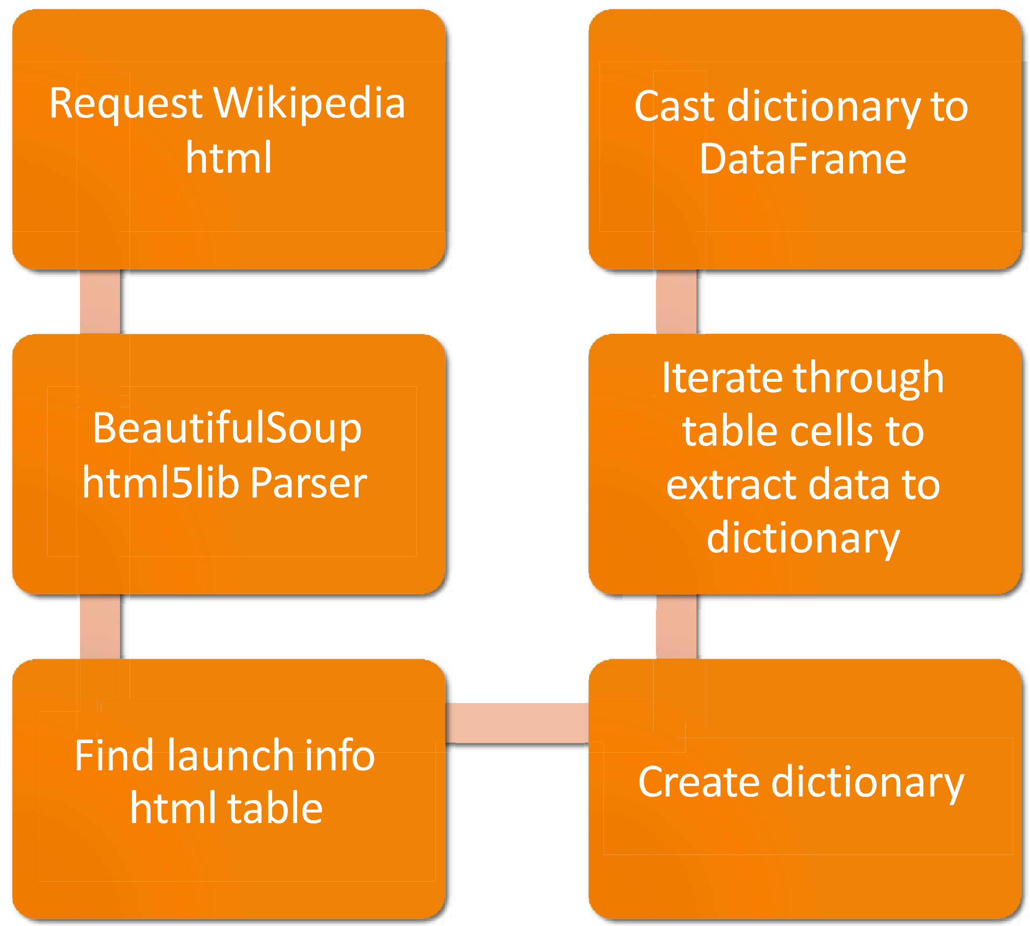
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Data Collection API

Data Collection – Web Scraping



Web Scraping

Data Wrangling

Create a training label with landing outcomes where successful = 1 & failure = 0.

The outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

Data Wrangling

EDA with SQL

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various payload sizes of customers and booster versions, and landing outcomes.

EDA with SQL

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend.

Scatter plots, line charts, and bar plots used to compare relationships between variables to decide if a relationship exists to utilize for training the machine learning model.

[Data Visualization]

Build an Interactive Map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

Utilized to understand why launch sites may be located where they are and visualize successful landings relative to location.

[Interactive Map]

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

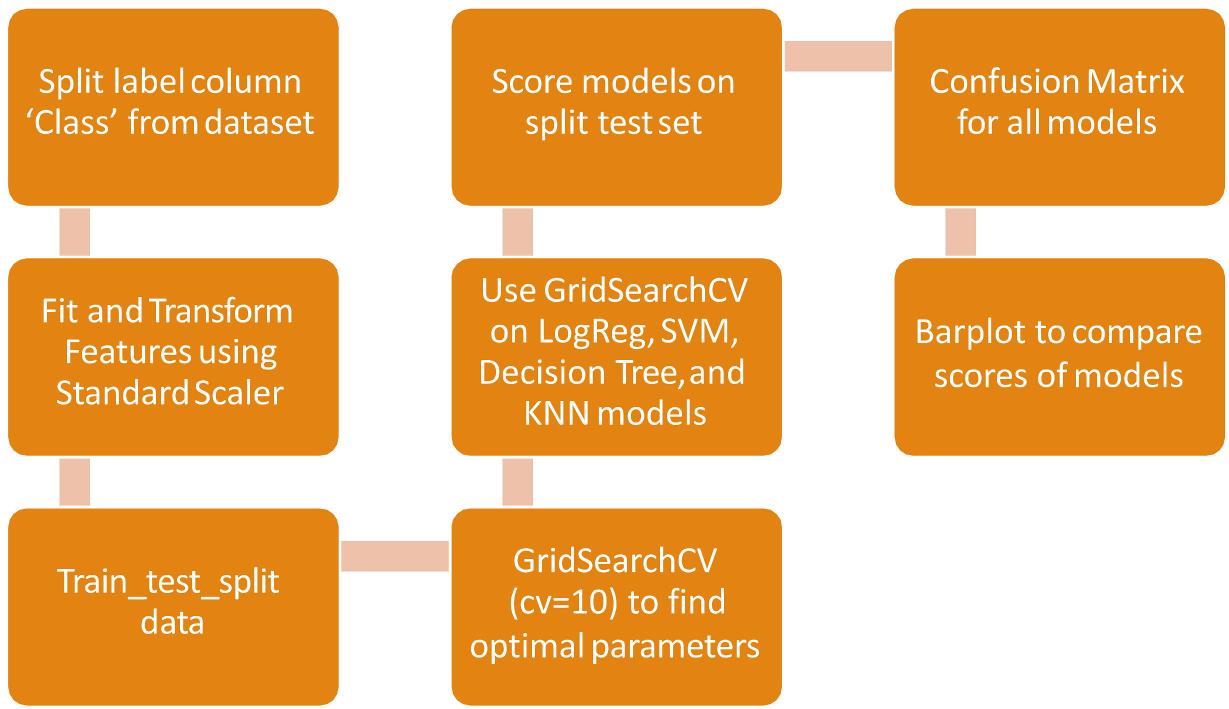
Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and booster version categories.

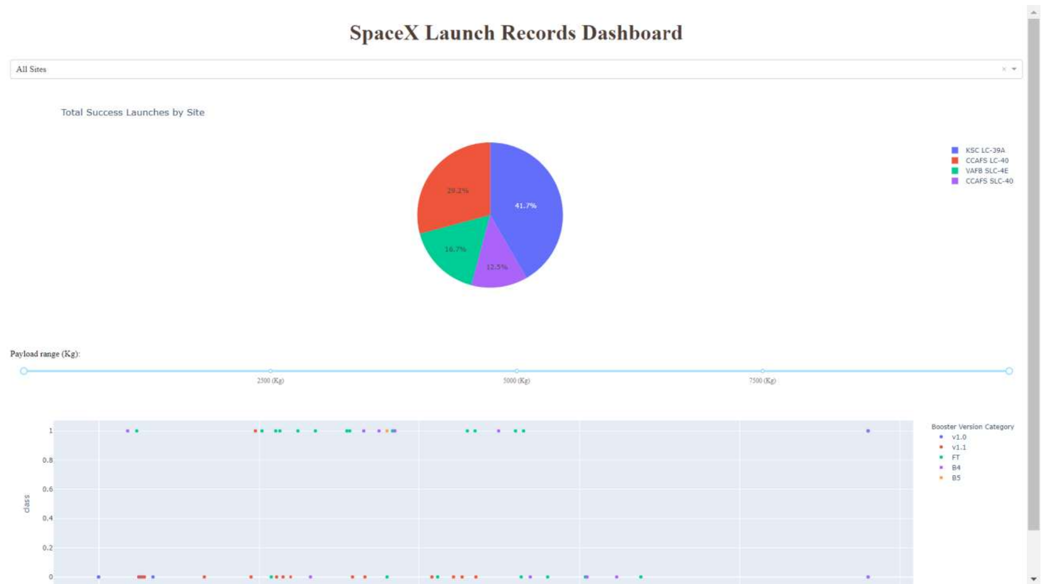
Dashboard

Predictive Analysis (Classification)



Machine Learning

Results

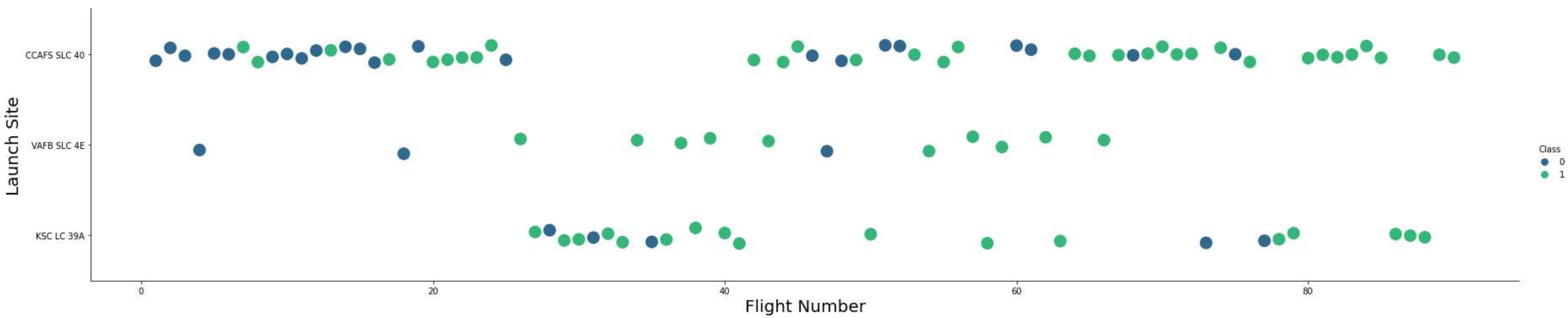


The above preview represents the Plotly dashboard, showcasing EDA with visualization, EDA using SQL, an interactive map with Folium, and the model's results achieving 83% accuracy.

EDA with Visualization

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

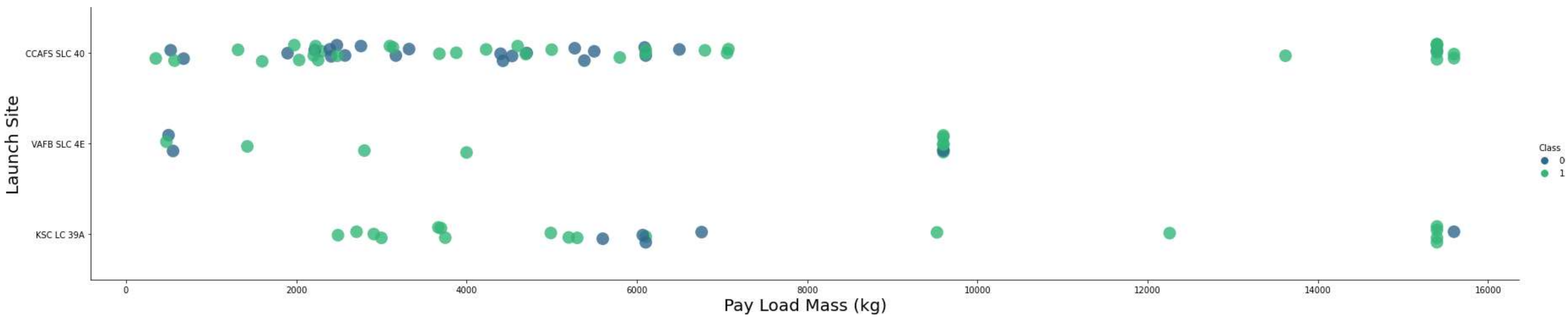
Flight Number vs. Launch Site



* Green indicates successful launch; Purple indicates unsuccessful launch

- Graphic indicates an increase in the success rate over time, as shown by the Flight Numbers
- There is a significant breakthrough around flight 20, significantly improving the success rate
- Cape Canaveral Air Force Station (CCAFS) is the primary launch site with the highest volume of flights

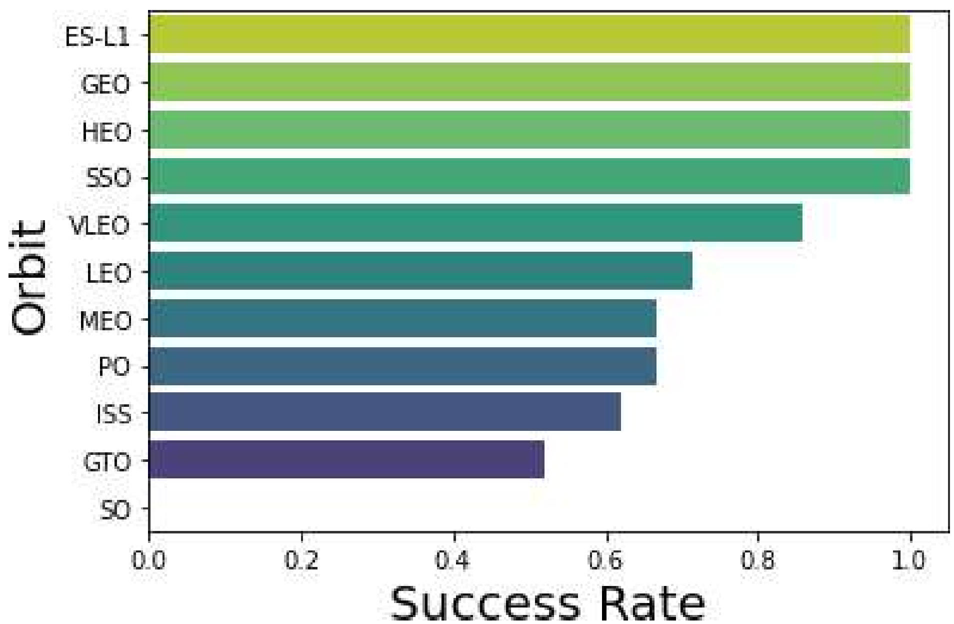
Payload vs. Launch Site



* Green indicates successful launch; Purple indicates unsuccessful launch

- Payload mass appears to be between 0-6000 kg.
- Different launch sites also seem to use different payload masses

Success Rate vs. Orbit Type



Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

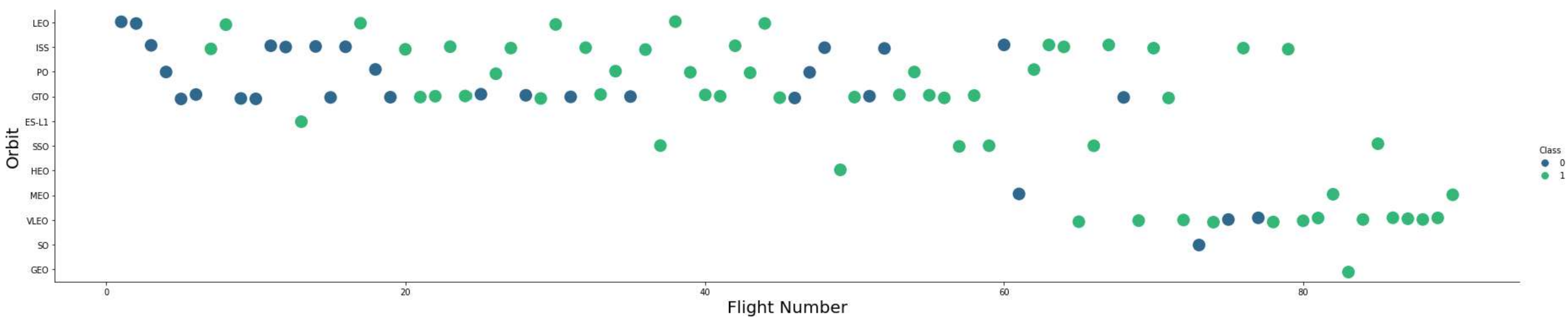
ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate, but has the largest sample

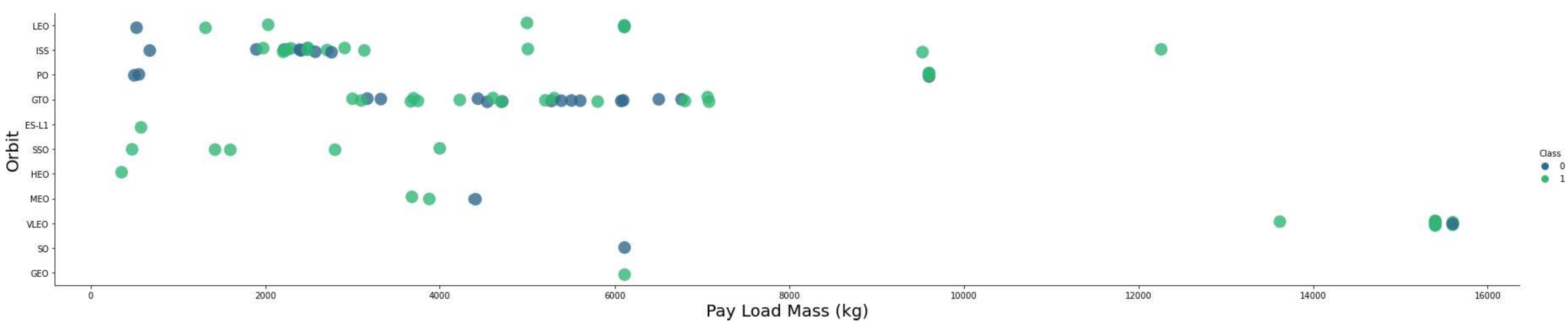
Flight Number vs. Orbit Type



* Green indicates successful launch; Purple indicates unsuccessful launch

- Launch orbit preferences have changed based on flight number and launch outcomes
- These changes seem to correlate with shifting preferences
- SpaceX initially focused on low Earth orbits, which experienced moderate success.
- Recently, they have returned to very low Earth orbit (VLEO) for their latest launches
- Overall, SpaceX appears to perform better in lower orbits.

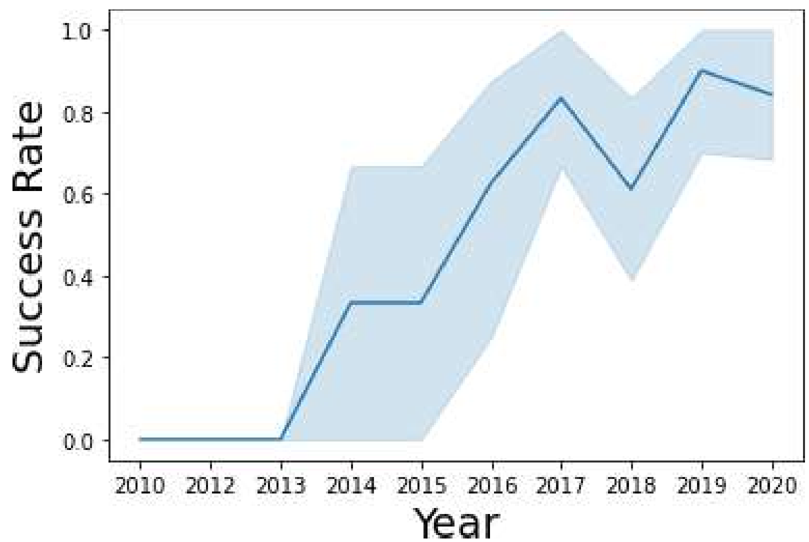
Payload vs. Orbit Type



* Green indicates successful launch; Purple indicates unsuccessful launch

- Payload mass appears to correlate with the type of orbit
- Low Earth Orbit (LEO) and Sun-Synchronous Orbit (SSO) have relatively low payload masses
- Conversely, the very Low Earth Orbit (VLEO) only shows payload mass values at the higher end of the range

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

- Success generally increased over time since 2013, with a slight dip in 2018
- Success in recent years at around 80%

EDA with SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

All Launch Site Names

```
In [4]: %%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;

 * ibm_db_sa://ftb12020:***@0c77d6f
Done.
Out[4]:
```

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
launch site with data entry errors.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Beginning with `CCA`

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The first five entries in the database with the Launch Site name begin with CCA.

Total Payload Mass from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services, which indicates that these payloads were sent to the International Space Station.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

This query calculates the average payload mass or launches that used booster version F9 v1.1.

The average payload mass of F9 1.1 is on the low end of our payload mass range.

First Successful Ground Pad Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

This query returns the first successful ground pad landing date.

The first ground pad landing wasn’t
until the end of 2015.

Successful landings in general appear starting 2014.

Successful Drone Ship Landing with Payload Between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

This query returns the four booster versions with successful drone ship landings and a payload mass between 4000 and 6000 non-inclusive.

Total Number of Each Mission Outcome

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status, and unfortunately, one failed in flight.

Boosters that Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

This query returns the booster versions that carried the highest payload mass of 15600 kg.

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

2015 Failed Drone Ship Landing Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

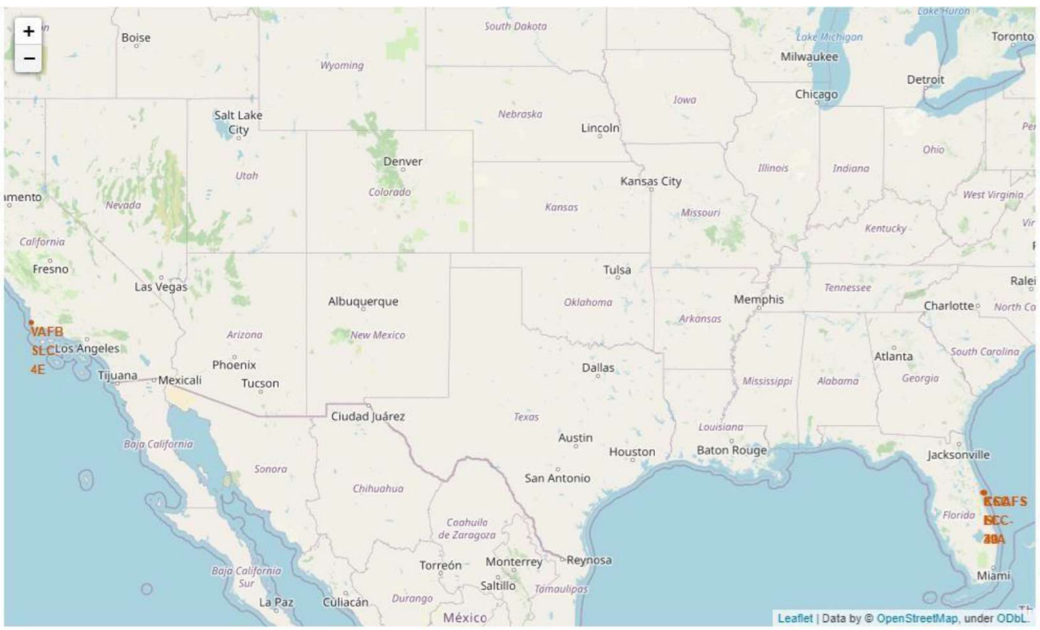
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

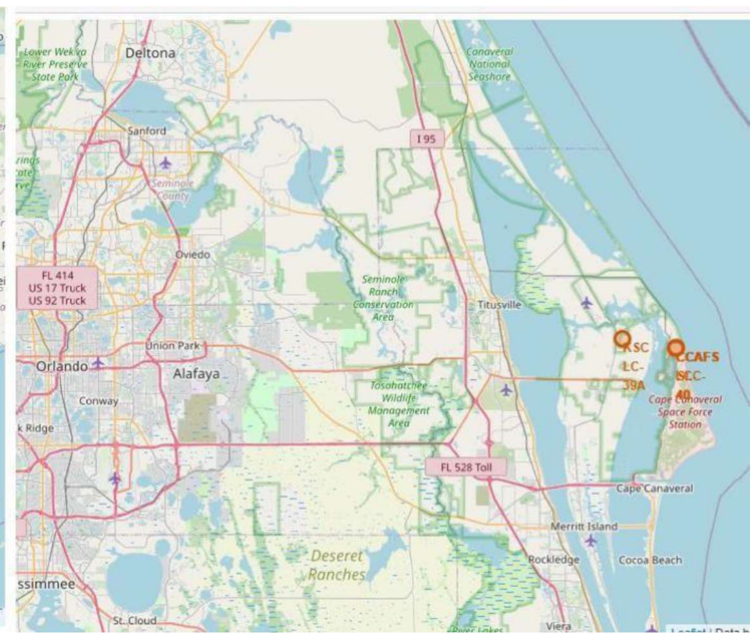
There are two types of successful landing outcomes: drone ship and ground pad landings.

There were 8 successful landings in total during this time period.

Interactive Map with Folium

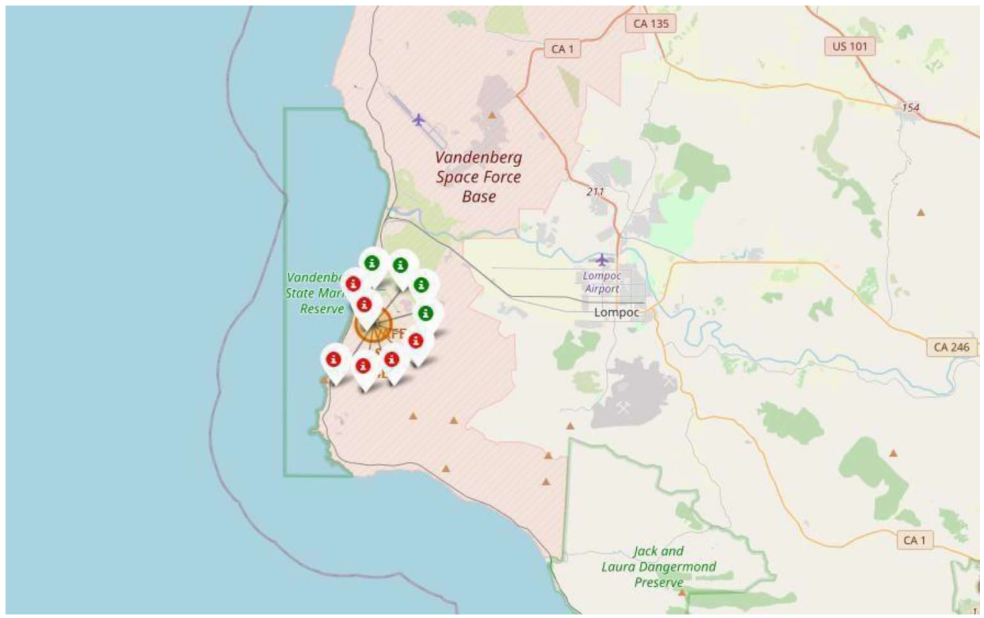
Launch Site Locations





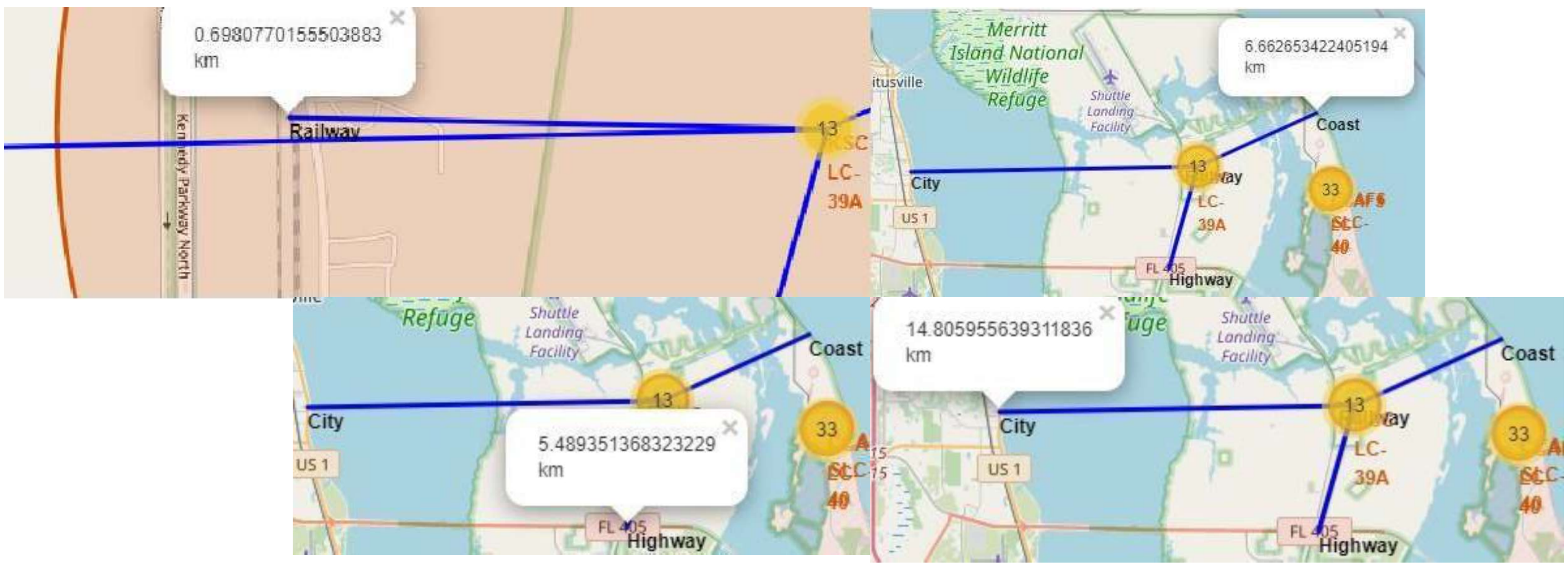
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

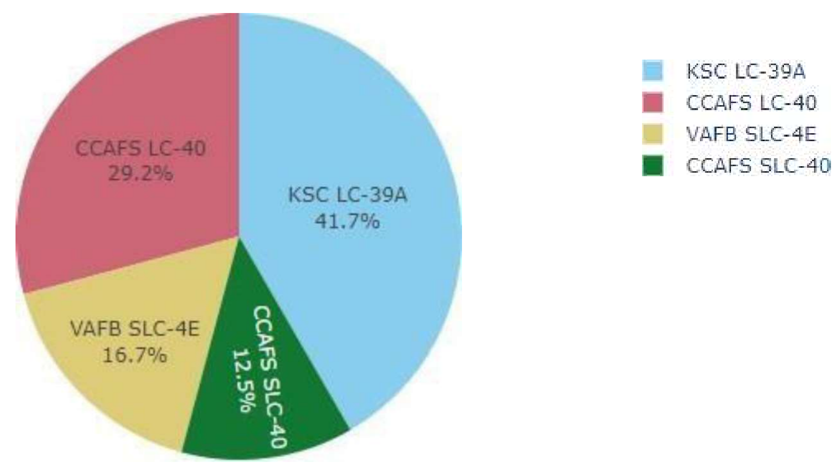
Key Location Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large parts and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling in densely populated areas.

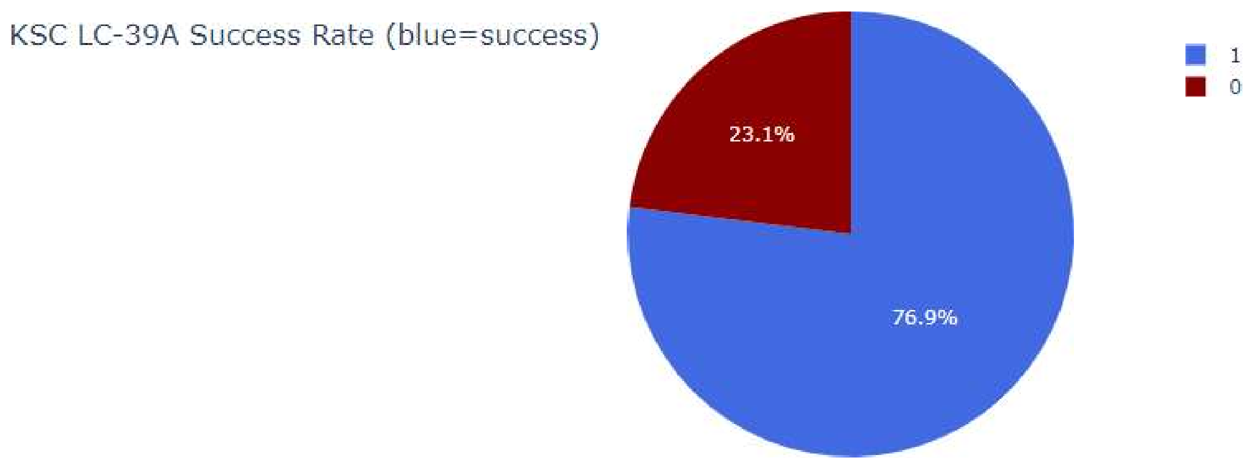
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



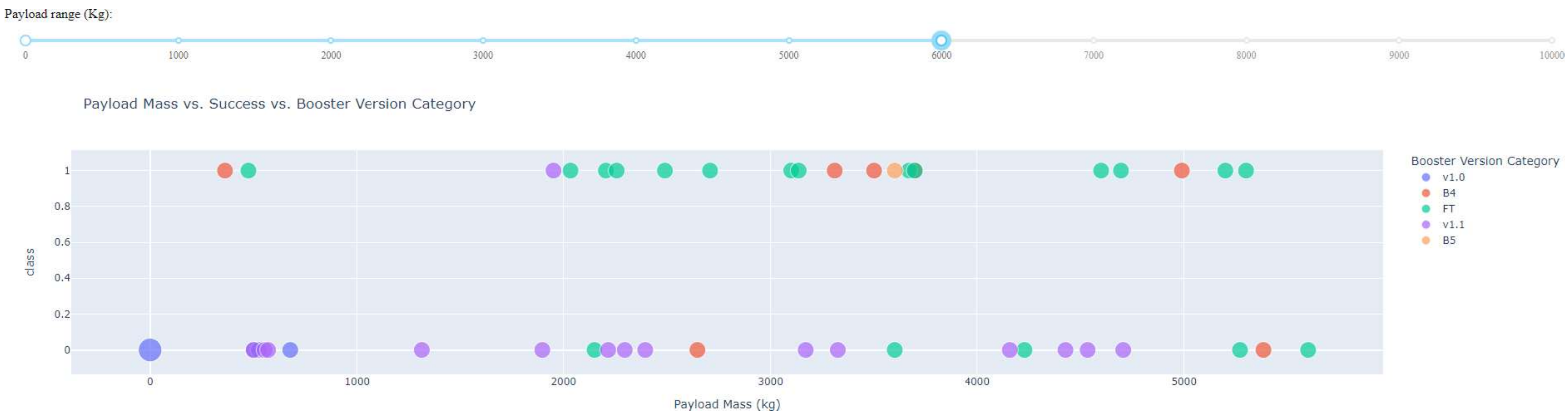
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40, so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings were performed before the name change. VAFB has the smallest share of successful landings. This may be due to the smaller sample and the increase in the difficulty of launching on the West Coast.

Highest Success Rate Launch Site



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Payload Mass vs. Success vs. Booster Version Category

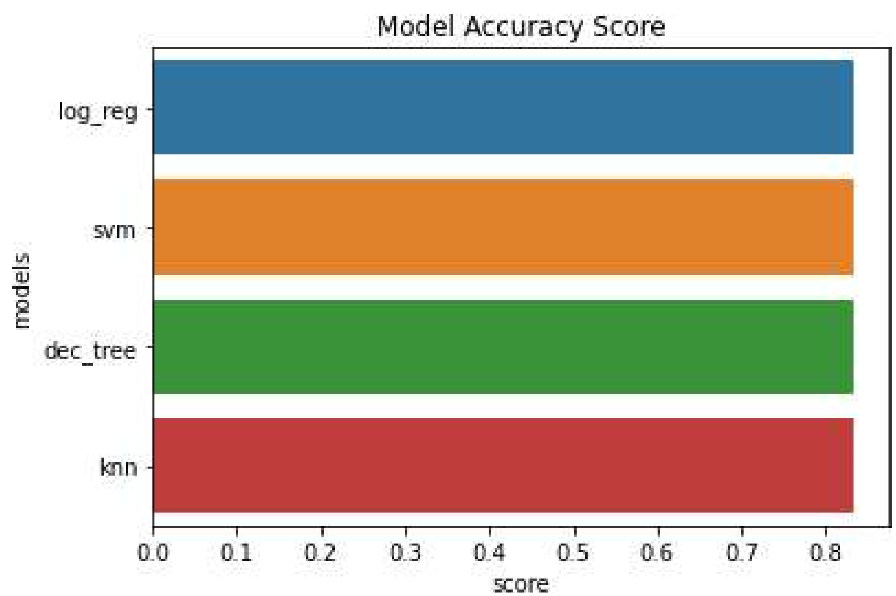


Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Predictive Analysis (Classification)

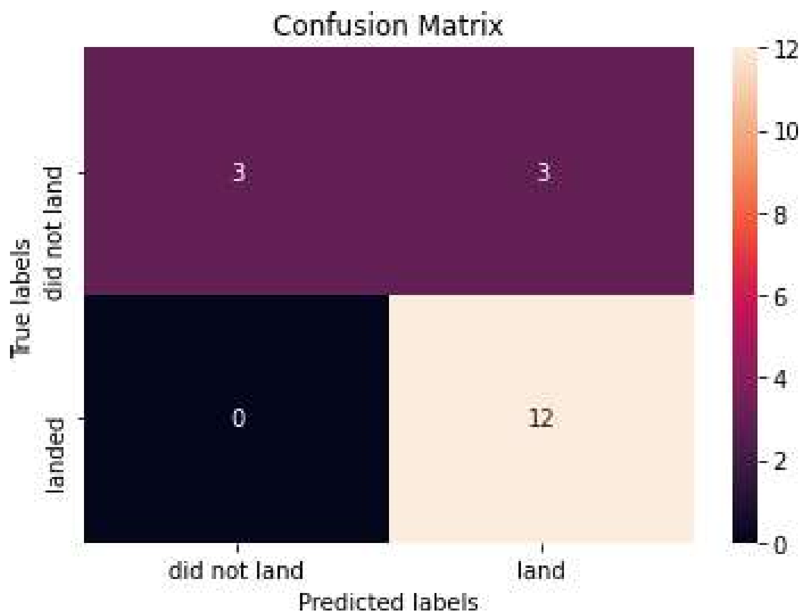
GRIDSEARCHCV(CV=10) ON LOGISTIC REGRESSION, SVM, DECISION TREE, AND KNN

Classification Accuracy



- All models had virtually the same accuracy on the test set at 83.33% accuracy.
  It should be noted that the test size is small at only a sample size of 18
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs
- Likely need more data to determine the best model

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

- Since all models performed the same for the test set, the confusion matrix is the same across all models
- Models predicted 12 successful landings when the true label was successful landing
- Models predicted 3 unsuccessful landings when the true label was unsuccessful landing
- Models predicted 3 successful landings when the true label was unsuccessful landings (false positives)
- Models overpredict successful landings

CONCLUSION

- **Task: To develop a machine learning model for SpaceY which wants to bid against SpaceX**
- **The goal of the model is to predict when Stage 1 will successfully land to save ~$100 million**
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data in a DB2 SQL database
- Created a dashboard for visualization
- Created a machine learning model with an accuracy of 83%
- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible, more data should be collected to better determine the best machine learning model and improve accuracy